Supporting Information (SI Appendix)

Riches of Phenotype Computationally Extracted from Microbial Colonies

Tzu-Yu Liu[1,2,]*, Anne E. Dodson[3,]*, Jonathan Terhorst[4], Yun S. Song[1,2,4,5, +] and Jasper Rine[3, +]

[1] Department of Mathematics and Department of Biology, University of Pennsylvania, PA 19104
[2] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720
[3] Department of Molecular and Cell Biology and California Institute for Quantitative Biosciences, University of California, Berkeley, CA 94720
[4] Department of Statistics, University of California, Berkeley, CA 94720
[5] Department of Integrative Biology, University of California, Berkeley, CA 94720
*These authors contributed equally to this work
[+]To whom correspondence should be addressed

Supplemental Figures and Text

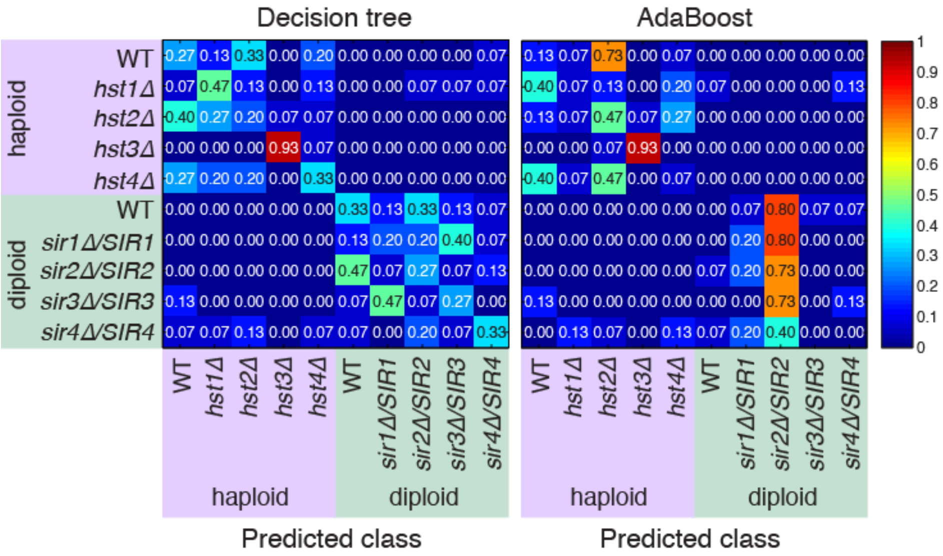


Figure S1. Confusion matrices by decision tree and AdaBoost on the multi-class classification of wild type and mutants, including both the haploid and diploid strains. Each row of the confusion matrix represents a different genotype (actual class), and the values within a row show the proportion of colonies that were predicted by the classifier to belong to the genotype specified by each column (predicted class). The color intensity, ranging from 0 to 1, corresponds to the fraction of colonies that were assigned to a particular predicted class. Successful classification results in high values along the diagonal, where each actual genotype intersects with its corresponding predicted genotype.

Image segmentation

Image segmentation plays an important role in modern biomedical imaging applications. Common methods include intensity thresholding using global or local threshold methods (1, 2), feature detection including edge detection (3) or Gabor filter (4), morphological filtering (5), region growing (6), classification or clustering (7, 8), and deformable models (9), etc. More recent developments include normalized cuts that build hierarchical partitions by using the low-level coherence (10), graph based approaches using pairwise region comparison (11), and energy minimization (12), etc. We refer the reader to some general surveys in the literature (13, 14). Here, we applied these recent segmentation techniques including normalized cuts (10) and energy minimization (12) to the collected images to segment the colonies grown on the petri dish, shown in Figure S2. However, these solutions were inadequate. We implemented the colony segmentation method described in the main text and illustrated in Figure S3.

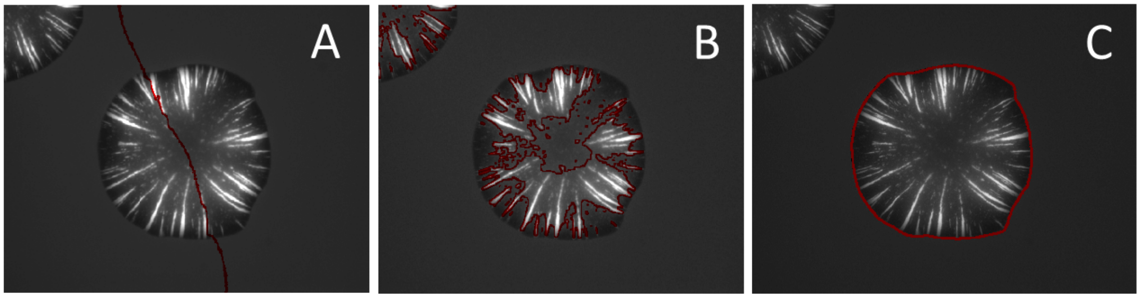


Figure S2. Image segmentation comparison on raw images. Image segmentation by **(A)** normalized cuts (10), **(B)** energy minimization (12) and **(C)** the method described in Figure S3.

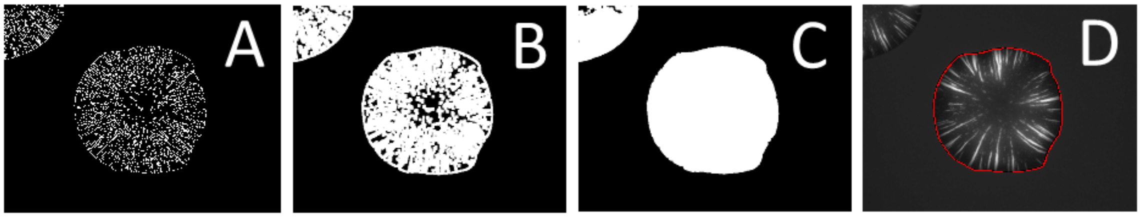


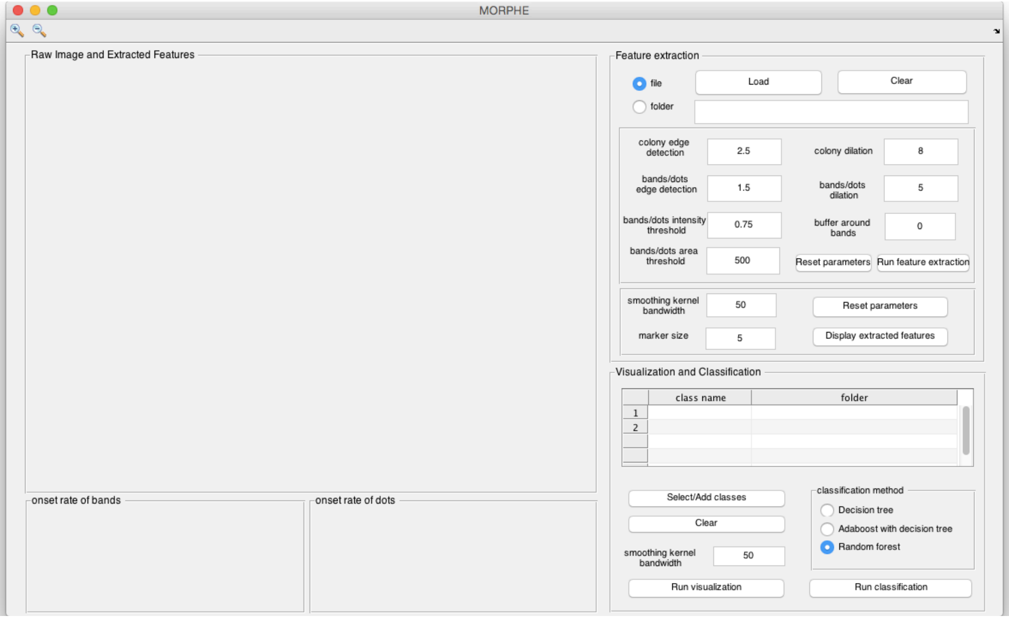
Figure S3. Segmentation of the colonies. **(A)** Edge detection was applied to the raw image. **(B)** The detected edges were dilated to form a closed boundary surrounding the colony of interest. **(C)** The interior pixels of the colonies were detected. **(D)** Identification of the colony with the largest area by counting the number of connected components found in (C) and the area of each connected component. The boundary of the colony of interest is outlined with the red curve.

Ensemble learning methods

AdaBoost was designed for binary classification. It repeatedly applies a base learning algorithm while maintaining a set of weights (15, 16). The weights are uniformly distributed at the initiation. Then at each round, the weights are updated such that the misclassified samples have larger weights. Let $w_{t,i}$ be the weight on sample i at the t round, and let ϵ_t be the error rate measured with respect to the distribution $w_{t,i}$ using the classifier $h_t(\boldsymbol{x}_i)$. The classifier $h_t(\boldsymbol{x}_i)$ is a weak learner, which can be a decision tree for instance. Then update the weights $w_{t+1,i} = w_{t,i}e^{\alpha_t I(y_i \neq h_t(x_i))}$, in which $\alpha_t = \log((1-\epsilon_t)/\epsilon_t)$, I is the indicator function, and $h_t(\boldsymbol{x}_i)$ is the classifier prediction for sample $\boldsymbol{x}_i$. AdaBoost.M2 generalizes the method to multi-class classification problems by reducing the multiclass problem to a larger binary problem (15). Random forest can be viewed as a bagging method, which is combined with random feature selection. Each tree is trained independently, in which it depends on the values of a random sample. Suppose we sample with replacement the rows of $\boldsymbol{X}$ and $\boldsymbol{Y}$ for N times to obtain a bootstrap sample $\boldsymbol{X}^b$ and $\boldsymbol{Y}^b$, where $b = 1,2,\dots,B$. Then a decision tree is trained with random selection of features at each node to determine the split. The bagging prediction is defined by taking the average among these trees (17, 18). We refer the reader to (19, 20) for discussion of the success of these ensemble learning methods.

Software

We have implemented the proposed algorithm with a user-friendly graphical interface. The software is implemented using MATLAB. Figure S4 shows the graphical interface. The software consists of mainly two sections. First, the feature extraction section enables the user to load an image or a batch of images, and set the parameters. All the parameters shown in Figure S4 are the default settings, and have been used for all the analysis presented in this paper. The second part on visualization and classification enables the user to generate heatmaps as shown in Figure 3 and Figure 4, along with the boxplots of the onset of bands and dots of each class. These extracted features can be used to differentiate the classes specified by the user. Three classification methods presented in the paper were incorporated into the software: decision tree, AdaBoost, and random forest. One can also use the software as a feature extraction and visualization toolbox and run further analysis using the extracted information. To use the software, create a folder for the colony images of the same class, and create two subfolders "GFP" and "RFP". The GFP folder should contain the images of GFP fluorescence, each ends in "_c1.tif". The RFP folder should contain the images of RFP fluorescence, each ends in "_c2.tif" and corresponds to a GFP image with the same prefix. A folder named "statistics" will be created after the analysis, in which the extracted features in .mat format and images of the features overlaid on top of the raw images are saved.



● file
○ folder

colony edge detection 2.5
colony dilation 8
bands/dots edge detection 1.5
bands/dots dilation 5
bands/dots intensity threshold 0.75
buffer around bands 0
bands/dots area threshold 500

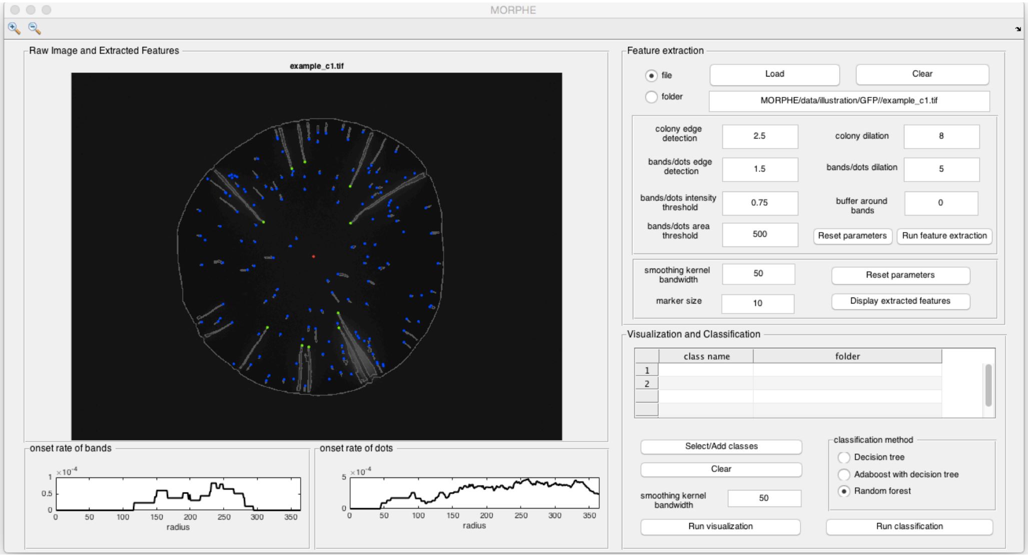
smoothing kernel bandwidth 50
marker size 5

	class name	folder
1		
2		

smoothing kernel bandwidth 50

classification method
○ Decision tree
○ Adaboost with decision tree
● Random forest

Figure S4. The graphical interface consists of two parts: 1) the feature extraction and 2) visualization and classification.



● file
○ folder
MORPHE/data/illustration/GFP//example_c1.tif

colony edge detection: 2.5
colony dilation: 8
bands/dots edge detection: 1.5
bands/dots dilation: 5
bands/dots intensity threshold: 0.75
buffer around bands: 0
bands/dots area threshold: 500

smoothing kernel bandwidth: 50
marker size: 10

smoothing kernel bandwidth: 50

classification method
○ Decision tree
○ Adaboost with decision tree
● Random forest

Figure S5. An example illustrating the extracted features overlaid on the raw image, and the extracted onset frequency of bands and dots over the colony.

The function of each parameter in the feature extraction panel is explained below.

(a) file/folder: Select file to analyze a single image; select folder for batch processing.
(b) Load: A dialog window opens for selecting the file or folder, depending on the specification of file/folder.
(c) Clear: Clear the selection. Return to step (a-b)
(d) colony edge detection: Sensitivity thresholds for the Canny edge detection method. The higher the threshold, the fewer colony edge pixels detected.
(e) colony dilation: The radius (pixels) of a flat, disk-shaped structuring element. The software uses the structuring element to dilate the binary image of (e), such that the detected edges form a close boundary of the colony.
(f) bands/dots edge detection: Sensitivity thresholds for the Canny method. The higher the threshold, the fewer edge pixels of bands/dots detected.
(g) bands/dots dilation: The radius (pixels) of a flat, disk-shaped structuring element. The software uses the structuring element to dilate the binary image of (f), such that the detected edges form a close boundary of each band/dot.
(h) bands/dots intensity threshold: A number between 0 and 1, denoted as α. A local threshold is set as the αth quantile of the pixel intensities of each connected component formed by detected bands/dots in (f-g). Each local threshold is applied to the interior pixels of the corresponding connected component. If the intensity of an interior pixel is above than the threshold, it is labeled as bands/dots.

(i) buffer around bands: The width (pixels) of a buffer around bands to be excluded from analysis. This is useful when there exist large bands with high fluorescence intensity that may create bias in the estimation.
(j) bands/dots area threshold: The threshold used to divide the detected bands/dots into the class of bands and the class of dots. If the number of pixels of a connected component is above the threshold, that connected component is labeled as a band; otherwise, it is labeled as a dot.
(k) Run feature extraction: Apply the parameters specified in step (d-j) to extract the features. To use the default settings, press the button “Reset parameters”.
(l) smoothing kernel bandwidth: The kernel bandwidth applied to the onset frequency of bands and the onset frequency of dots, illustrated in Figure 2.
(m) marker size: the marker size of the green and blue dots shown in Figure S5.
(n) Display extracted features: Display the extracted features overlaid on top of the raw images, as shown in Figure S5.
(o) To use the default settings, press the button “Reset parameters”.
(p) To analyze another file or batch, return to step (a).

The function of each parameter in the visualization and classification panel is explained below.
(a) Select/Add classes: A dialog window opens for selecting a folder to classify. Edit the class name of each folder. Folders with the same class name will be treated as the same class. There must be at least two classes specified.
(b) Clear: Clear the selection in step (a). Return to step (a).
(c) smoothing kernel bandwidth: the kernel bandwidth applied to the onset rate of bands and the onset rate of dots.
(d) Run visualization: Create heatmaps of the smoothed and unsmoothed onset rate of bands and the onset rate of dots, as shown in Figure 3.
(e) Decision tree / Adaboost with decision tree / Random forest: Select the classification method.
(f) Run classification: Use the features smoothed by the kernel bandwidth specified in step (c) to predict the class labels. A heatmap of the confusion matrix will appear after leave-one-out test is done.

Supplemental Tables

Yeast strains

Strain	Genotype
JRY9628	matΔ::natMX lys2 his3-11,15 leu2-3,112 can1-100 HMLα-α2Δ::cre ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-kanMX-loxP-yEGFP-T_{ADH1}
JRY9634	matΔ::natMX lys2 his3-11,15 leu2-3,112 can1-100 HMLα-α2Δ::cre ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-kanMX-loxP-yEGFP-T_{ADH1} hst1Δ::hphMX
JRY9635	matΔ::natMX lys2 his3-11,15 leu2-3,112 can1-100 HMLα-α2Δ::cre ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-kanMX-loxP-yEGFP-T_{ADH1} hst2Δ::hphMX
JRY9636	matΔ::natMX lys2 his3-11,15 leu2-3,112 can1-100 HMLα-α2Δ::cre ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-kanMX-loxP-yEGFP-T_{ADH1} hst3Δ::hphMX
JRY9637	matΔ::natMX lys2 his3-11,15 leu2-3,112 can1-100 HMLα-α2Δ::cre ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-kanMX-loxP-yEGFP-T_{ADH1} hst4Δ::hphMX
JRY9731	MATα/matΔ::kanMX lys2/lys2 his3-11,15/his3-11,15 leu2-3,112/leu2-3,112 can1-100/can1-100 HMLα/HMLα-α2Δ::cre ura3-1/ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-kanMX-loxP-yEGFP-T_{ADH1}
JRY9732	MATα/matΔ::kanMX lys2/lys2 his3-11,15/his3-11,15 leu2-3,112/leu2-3,112 can1-100/can1-100 trp1-1/TRP1 sir1Δ::TRP1/SIR1 HMLα/HMLα-α2Δ::cre ura3-1/ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-kanMX-loxP-yEGFP-T_{ADH1}
JRY9733	MATα/matΔ::kanMX lys2/lys2 his3-11,15/his3-11,15 leu2-3,112/leu2-3,112 can1-100/can1-100 trp1-1/TRP1 sir2Δ::TRP1/SIR2 HMLα/HMLα-α2Δ::cre ura3-1/ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-kanMX-loxP-yEGFP-T_{ADH1}
JRY9734	MATα/matΔ::kanMX lys2/lys2 his3-11,15/his3-11,15 leu2-3,112/leu2-3,112 can1-100/can1-100 trp1-1/TRP1 sir3Δ::TRP1/SIR3 HMLα/HMLα-α2Δ::cre ura3-1/ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-kanMX-loxP-yEGFP-T_{ADH1}
JRY9735	MATα/matΔ::kanMX lys2/lys2 his3-11,15/his3-11,15 leu2-3,112/leu2-3,112 can1-100/can1-100 trp1-1/TRP1 sir4Δ::TRP1/SIR4 HMLα/HMLα-α2Δ::cre ura3-1/ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-kanMX-loxP-yEGFP-T_{ADH1}
JRY10639	MATα/matΔ::natMX lys2/lys2 his3-11,15/his3-11,15 leu2-3,112/leu2-3,112 can1-100/can1-100 HMLα/HMLα-α2Δ::cre ura3-1/ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-hphMX-loxP-yEGFP-T_{ADH1}
JRY10640	MATα/matΔ::natMX lys2/lys2 his3-11,15/his3-11,15 leu2-3,112/leu2-3,112 can1-100/can1-100 HMLα/HMLα-α2Δ::cre ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-kanMX-loxP-yEGFP-T_{ADH1}/ ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-hphMX-loxP-yEGFP-T_{ADH1}
JRY10641	MATα/matΔ::natMX lys2/lys2 his3-11,15/his3-11,15 leu2-3,112/leu2-3,112 can1-100/can1-100 HMLα-α2Δ::cre/HMLα-α2Δ::cre ura3-1/ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-hphMX-loxP-yEGFP-T_{ADH1}
JRY10642	MATα/matΔ::natMX lys2/lys2 his3-11,15/his3-11,15 leu2-3,112/leu2-3,112 can1-100/can1-100 HMLα-α2Δ::cre/HMLα-α2Δ::cre ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-kanMX-loxP-yEGFP-T_{ADH1}/ ura3Δ::P_{GPD}-loxP-yEmRFP-T_{CYC1}-hphMX-loxP-yEGFP-T_{ADH1}

Table S1. All strains used in this study were derived from W303.

Significance test of the onset frequencies of various haploid and diploid yeast strains

Strain	Relevant genotype	Avg onset frequency	Two-sample t-test with WT
JRY9628	WT haploid	8.33e-04	
JRY9364	*hst1Δ*	4.33e-04	8.13e-06
JRY9635	*hst2Δ*	6.47e-04	4.55e-02
JRY9636	*hst3Δ*	2.07e-03	8.15e-10
JRY9637	*hst4Δ*	7.31e-04	2.64e-01

Table S2. Average of the mean onset frequencies and significance test of haploid cells. The mean onset frequencies were averaged over colonies of the same class.

Strain	Relevant genotype	Avg onset frequency	Two-sample t-test with WT
JRY9731	WT diploid	8.74e-05	
JRY9732	*sir1Δ/SIR1*	1.42e-04	1.33e-02
JRY9733	*sir2Δ/SIR2*	7.37e-05	3.38e-01
JRY9734	*sir3Δ/SIR3*	1.99e-04	6.11e-04
JRY9735	*sir4Δ/SIR4*	2.74e-04	6.02e-02

Table S3. Average of the mean onset frequencies and significance test of diploid cells. The mean onset frequencies were averaged over colonies of the same class.

Strain	Ploidy	*HML::cre*	RFP-GFP	Avg onset frequency	Two-sample t-test with JRY10639
JRY9628	1n	1	1	7.66e-04	6.24e-14
JRY10639	2n	1	1	1.23e-04	
JRY10640	2n	1	2	3.21e-04	2.19e-05
JRY10641	2n	2	1	3.92e-04	5.54e-06
JRY10642	2n	2	2	7.33e-04	1.82e-10

Table S4. Average of the mean onset frequencies and significance test of cells with the indicated copy numbers of *HML::cre* and the RFP-GFP cassette. The mean onset frequencies were averaged over colonies of the same class.

Significance test of the onset frequencies of colonies grown under various environmental conditions

[vitamin C]	Avg onset frequency	Two-sample t-test with 0mM
0 mM	10.00e-04	
0.1 mM	9.48e-04	3.29e-01
1 mM	8.89e-04	2.07e-02
10 mM	7.26e-04	7.42e-06

Table S5. Average of the mean onset frequencies and significance test of colonies grown under varying levels of vitamin C. The mean onset frequencies were averaged over colonies of the same class.

[$NiCl_2$]	Avg onset frequency	Two-sample t-test with 0mM
0 mM	9.15e-04	
0.05 mM	5.45e-04	2.16e-13
0.1 mM	3.98e-04	6.19e-17

Table S6. Average of the mean onset frequencies and significance test of colonies grown under varying levels of $NiCl_2$. The mean onset frequencies were averaged over colonies of the same class.

[H_2O_2]	Avg onset frequency	Two-sample t-test with 0mM
0 mM	1.14e-03	
0.1 mM	1.07e-03	1.71e-01
0.5 mM	9.32e-04	1.74e-04

Table S7. Average of the mean onset frequencies and significance test of colonies grown under varying levels of H_2O_2. The mean onset frequencies were averaged over colonies of the same class.

Sugar	Avg onset frequency	Two-sample t-test with Glucose
Glucose	8.81e-04	
Galactose	2.41e-03	1.80e-21
Raffinose	2.89e-03	8.84e-26

Table S8. Average of the mean onset frequencies and significance test of colonies grown with different sugars. The mean onset frequencies were averaged over colonies of the same class.

References

1. Polakowski WE, et al. (1997) Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency. *IEEE Trans Med Imaging* 16(6):811–819.
2. Wu K, Gauthier D, Levine MD (1995) Live cell image segmentation. *IEEE Trans Biomed Eng* 42(1):1–12.
3. Senthilkumaran N, Rajesh R (2009) Image segmentation - A survey of soft computing approaches. *ARTCom 2009 - International Conference on Advances in Recent Technologies in Communication and Computing*, pp 844–846.
4. Jain AK, Farrokhnia F (1991) Unsupervised texture segmentation using Gabor filters. *Pattern Recognit* 24(12):1167–1186.
5. Meyer F, Beucher S (1990) Morphological segmentation. *J Vis Commun Image Represent* 1(1):21–46.
6. Pohlman S, Powell KA, Obuchowski NA, Chilcote WA, Grundfest-Broniatowski S (1996) Quantitative classification of breast tumors in digitized mammograms. *Med Phys* 23(8):1337–1345.
7. Kapur T, Grimson WEL, Kikinis R, Wells WM (1998) Enhanced Spatial Priors for Segmentation of Magnetic Resonance Imagery. *MICCAI '98 Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), pp 457–468.
8. Coleman GB, Andrews HC (1979) Image segmentation by clustering. *Proceedings of the IEEE*, pp 773–785.
9. Davatzikos C (1996) Using a deformable surface model to obtain a shape representation of the cortex. *IEEE Trans Med Imaging* 15(6):785–795.
10. Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 22(8):888–905.
11. Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. *Int J Comput Vis* 59(2):167–181.
12. Delong A, Osokin A, Isack HN, Boykov Y (2012) Fast approximate energy minimization with label costs. *Int J Comput Vis* 96(1):1–27.
13. Meijering E (2012) Cell Segmentation: 50 Years Down the Road. *IEEE Signal Process Mag* 29(5):140–145.
14. Pham DL, Xu C, Prince JL (2000) Current methods in medical image segmentation. *Annu Rev Biomed Eng* 2:315–37.
15. Freund Y, Schapire RE (1997) A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. *J Comput Syst Sci* 55(1):119–139.
16. Freund Y, Schapire R (1999) A short introduction to boosting. *IJCAI International Joint*

Conference on Artificial Intelligence, pp 1401–1406.

17. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32.

18. Hastie T, Tibshirani R, Friedman J (2003) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer) doi:citeulike-article-id:161814.

19. Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: A statistical view of boosting. *Ann Stat* 28(2):337–407.

20. Wyner AJ, Olson M, Bleich J (2015) Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers. *arXiv Prepr*:1–40.